An Architecture to Enhance Post Retrieval Document Relevancy Using Integrated Techniques

Vimala Balakrishnan and Kian Ahmadi

Abstract—This paper is a proposal based on an on-going work in which we attempt to take advantage of Information retrieval (IR) and case-based reasoning (CBR) techniques combined with the aim to improve the document relevancy of a search result. The proposed architecture contains two main phases: first is the IR phase whereby relevance feedback (RF) is implemented on search results produced based on adjacency keyword algorithm. Second is the CBR phase which further improves the results based on the output from phase one. This paper presents an explanation on the proposed RF-CBR model. It is believed that the integration of these two popular techniques would result in an improved document relevancy.

Index Terms—Adjacency keywords, case-based reasoning, KNN algorithm, relevance feedback.

I. INTRODUCTION

With ever growing information over the Web, finding high quality relevant information within the large collection of texts is a challenging issue. Traditional searching for information relies on exact match or “one-fits-all” principle search mechanism based on the use of Boolean queries and keywords. Such an approach often results in the same documents being returned to the users whenever the same keywords (regardless of the order) are used in the queries. Additionally this also means an overwhelming number of results are returned to the users to select [1]. This prompted researchers to explore best match mechanism, which relies on unstructured queries and ranks the search documents based on their relevancy [2].

The case-based reasoning (CBR) is based on the best match mechanism. It works on the basis of suggesting or solving new problems by adapting previous solutions to the new problems. The CBR paradigm covers a range of different methods for organizing, retrieving, utilizing and indexing the knowledge retained in past cases and it favors learning from experience, since it is usually easier to learn by retaining a concrete problem solving experience [3]. CBR has been widely and successfully used in medical domains [4], however recent studies have attempted to utilize this technique in the field of information retrieval (IR) as well [5].

By considering the importance of document relevancy score in the ranked retrieval results and the advantages of CBR technique, this paper aims to propose an architecture that combines IR and CBR to improve document relevancy.

The two main objectives are to improve IR mechanisms by first, maximizing the relevancy of the information returned system outputs to those intended by the user(s), and second, enhancing the similarities of the information retrieval output of the system’s ranking to the human ranking using CBR.

The rest of the paper is organized as follows: the following section provides a brief description on IR. IR followed by previous works on relevance feedback (RF) and CBR. Then, the proposed architecture is presented and discussed. Future work and conclusion concludes the paper.

II. INFORMATION RETRIEVAL

Information retrieval (IR) is the study of techniques providing solution for retrieving text documents from relevant to the user needs. According to Manning et al. [6], IR is defined as “ways to find materials (usually documents) that contain large number of texts (of an unstructured that meet the information need from within large collections (usually stored on computers).” IR approach is to find items that are relevant and considered the best among the partial matching to the keywords defined query. It provides a solution with the best match results to rank to the degree of relevance in response to the user query [7]. A typical example of IR is the web search engine Google Search that is designed to search for information on the Web.

III. INFORMATION RETRIEVAL MODELS

There are three classic IR models: First, the retrieval model which is based on set theory and algebra that model documents and queries as sets of terms [8]. This is one of the earliest and simplest which keywords are logically combined with operators AND, OR and NOT to form the query in information retrieval system [8]. Although this retrieval approach is in many commercial systems, but the drawbacks are well-known. For example, based upon two Boolean “zero” and “one”, the retrieval results tend to have the large number of documents or none at all, and is difficult to control the output size [1]. Furthermore, retrieval return matching documents without relevance of the documents into consideration. This means the user having to browse through the list to find the one that meets his or her information’s need.

Second is the probabilistic retrieval model probability theory that ranks documents according the probability of relevance [9]. In other words, for a given